The Molecular Bases of Biology

Word-for-word transcript of Dr. Francis Crick's lecture at the CARLO ERBA FOUNDATION, January 7, 1963

It makes me very happy to be here to-day to give you this lecture, but when I look around and see so many of you I find it difficult to believe that you are all molecular biologists. Those of you who are molecular biologists will therefore have to excuse me if I speak in a rather general way about the molecular structure biology, and do not dwell too much on details and speak about things perhaps which are rather well known to you, so that the others of you may follow more easily what I have to say.

So let me begin by something which will be familiar to many of you, but perhaps will not be known to all of you. Let us look at nature as a whole; let us look at plants and at animals and microorganisms, viruses and so forth, the diseases which infect us — what do we first notice? We notice how different they are. Some are very big, some are very small; some of them are pink and some green, and so on. If you look at living things with the eyes of the lawyer or the engineer or something of this sort, you would not suspect that they have very much in common.

But when we look more closely, when we look at the biochemical level and see the molecules of which living things are made, a very striking thing is found, a thing which I think is not perhaps emphasized enough in the way we teach our students. We find that the *basic molecules* (which I will describe in a moment in more detail) *are astonishingly similar right throughout nature, from the very smallest to the very biggest, from the very simplest to the most complicated.*

And in particular we find, when we look at very simple living things, that there are two families of molecules — they are not simple molecules, they are quite complicated ones — there are two families which we find in almost all living things. And these families are firstly the nucleic acids, of which there are two sorts, DNA and RNA, and secondly the proteins. And the surprising thing is that the bits, *the small molecules of which the nucleic acids and the proteins are made are exactly the same throughout nature.* There are four different units which make up the nucleic acids, slightly different in the two but the same throughout nature; and there are twenty different units making up proteins, and they again are universal.

There is another property of the two families of molecules which is implied in what I said, namely that they are made in a simple, standard way. They are both long chain molecules; they both have a long chemical backbone; of course you know the backbone is different in the two cases but the backbone itself of the nucleic acids, the phosphate-sugar groups, or the backbone of the proteins, the polypeptide chains, is quite uniform as you go along from one end of the chain to the other. The difference is in the side groups that are attached, of which, as I said, there are four sorts in nucleic acids and 20 sorts in proteins. And we see here an enormous simplicity in the basic molecules of life. They are built on a simple plan, you might say, although they are big molecules they are not so very complicated molecules; they are built in a simple way, and they are uniform throughout nature.

But of course there are many other molecules of different sorts in living things, there are carbohydrates, and lipids, lots and small molecules — why therefore do I pick out these two families of molecules, proteins and nucleic acids? To answer this question we have to ask ourselves what are the most fundamental properties of living things. And in my mind, there is no doubt that the dominant characteristic of living organisms is firstly their complexity; and secondly, which is the most important thing, is the mechanism by which this complexity has arisen. *We believe that it has arisen by a random process, by changes which are not predirected* and a mechanism which enables you to produce great complexity from random events is that suggested more than a hundred years ago by Darwin: the mechanism of *natural selection*. And so we must ask ourselves: what is needed at molecular level to supply the molecules for natural selection?

One of the most fundamental properties is replication; you must have a situation such that one rather complicated thing can produce two like things, and then the two in turn can produce four, and they in turn produce eight: you must have geometrical replication. The second property is the opportunity of making alterations, or mutations, as we say, in the biological sense, in this message which is being passed on. It is not enough just to make a mistake; it must be a mistake of the type that can be copied — because mistake is the wrong word, it is an alteration, and some of these alterations may be advantageous. So the second property is — you must be able to make changes which themselves will be copied. This is not quite enough, because the organism has to live in a difficult and hostile environment and it must be able to deal chemically with all the processes and problems that it finds in that environment. In some way therefore an organism must be versatile: it must be able to cope with all the difficulties at chemical level and higher which it finds in its environment. So there must be three properties: replication, the opportunity for mutation, versatility.

Now, it is not obvious that these properties would find expression in the properties of molecules. But it happens that it is so. The interesting thing is that they are not expressed by one family of molecules, but by two. The nucleic acids are those which we associate with replication and mutation; but the property of versatility belongs to the proteins. And so I would speak briefly about the proteins first. Now, as we have said, proteins are made on a rather simple plan. *A typical polypeptide chain of a protein may contain from one to three hundred aminoacids, joined end to end,* of which, you remember, there are twenty different kinds. How then is it possible that such a molecule can act in a special way? So far we have only been discussing the chemical structure, the way the parts are poined together by strong chemical bonds. But when the molecule is synthesized, it folds up on itself by means of rather weak bonds and forms an´ intricate tridimensional structure, so that very elaborate chemical patterns may be made by means of a very simple basic plan of synthesis.

We then ask, what is it that the protein molecules do as a class, and the most important function that they have is to act as enzymes, as the catalysts which speed up the chemical reactions within living cells. Each type of enzyme typically catalyzes one particular chemical reaction. And a small bacterial cell, we believe, has probably several thousand different kinds of enzymes inside it. And each one in made in a specific way, has a specific pattern on its surface, and part of its surface acts as a catalytic centre. I think this property of proteins, that they are built in a simple way but behave in a complicated and subtle way, is one of the most important properties of living things. It is interesting to compare it to the way we write, the way we express ourselves. *If you look at our language, a paragraph in our twenty-odd-letter language is rather about the same length as a protein.* It is one long chain although we write it on several lines — in fact, the language is continuous. It is made of simple units put end to end. The meaning of the paragraph depends on the exact relation of the letters which make up the words and the arrangement of the words that make up the paragraph. We too us complicated and subtle ideas by means of a simple alphabet.

So we must ask what it is about proteins, that they cannot be the only family of molecules we find in living things. And the answer is a surprising one: it is that the polypeptide chain of the protein likes to fold up, to a first approximation, in a regular, simple helix on itself. It is not a very stable helix — it may go as a helix for part of the time and then turn a corner, and because it is not very stable therefore it is easy to bend it and make it fold back into this intricate type of pattern. What, however, this property means, or another property let us say, is that, however, you do not have a structure of protein made of two chains. You might say that the protein is rather like a cultivated bachelor who is very versatile and can do many things but does not like to settle down to married life.

So let us now look at the other family of molecules, the nucleic acids. Now what we find there is exactly the opposite. What we find is that in that case if you have a single chain of nucleic acid, of DNA or RNA, it does not take on a regular form in space; but it is very easy for nucleic acids to form a two-stranded (or even a three-stranded) structure like a twisted rope-ladder, as in the model, and it is this property which makes it so suitable for replication.

As the replication scheme will be familiar to many of you so I only describe it very briefly. As you know, there are four types of side groups stuck on here, two big ones and two small ones, and the copying is done by separating these two chains and arranging when you have a single chain that you build a second new companion chain on it — so the bits fit together. So if you have a big one on the old chain you must have a little one on the new chain. The little knobs, the little projections on the molecules are such that you can only have the right one fitting. If you have adenine here you must have thymine there; if you have guanine on this side you must have cytosine on the other. It is this that enables you to have diversity at the same time as a regular structural background.

So now we should ask: if nucleic acid is such a suitable molecule for replication, why cannot THAT be the basis for living things. But the trouble is the poor nucleic acid molecule can only do this one thing; it has only got four side groups instead of twenty, they are all rather alike; it cannot build, we believe, elaborate structures of this type, it must either be very disorganized or all too regular, and so it does not have the versatility which is needed in order to defend itself in a difficult and hostile world. If we have to make comparisons, we could say that *nucleic acid is like a married lady of the old school, good for re-*

production but not good for anything else.

So we see that we must have these two classes of molecules and now we must ask how they are related one to the other. We have seen the structure of the genetic material and the structure of the gene product, the protein. And we find there is an elaborate arrangement whereby these two are connected. In brief, *the sequence of the twenty sorts of aminoacids in the proteins is determined by the sequence in a certain par or stretch of the nucleic acid chain by the four types of things there.* In the conventional language we believe very roughly, in the old terminology, that you have one gene, one genetic unit, that may be perhaps a thousandth of these things long. One gene determines one protein. The new jargon for the benefit of the expert is a little more complicated, it is one cystron to one polypeptide chain. But the one-gene -one-protein is a very useful idea.

Now in order to do this there has to be a very elaborate biochemical machinery for protein synthesis. This is a little difficult to describe in a simple way. In brief, we believe that in most cells, certainly in simple cells, *the DNA is in the genes and the genes are in the nucleus.* But most of the synthesis of the proteins takes place outside the nucleus of the cell, in the outer part of the cell, in the cytoplasm, and it takes place on little particles which are not unlike viruses but are quite distinct viruses, called ribosomes. And in some way *the genetic message has to pass from the DNA in the nucleus to the ribosomes in the cytoplasm. We believe this is done by a special sort of RNA, the other nucleic acid, which is now called messenger RNA. The DNA is the permanent file copy.* In a simple cell there is just one copy in the cell. In man, as we are more complicated organisms, we usually have two copies one from our father, one from our mother. This is the original copy. *But the messenger RNA is the working copy which is sent out into the rest of the cell, and many copies of messenger RNA are made from one piece of DNA.* So that is the broad picture, a particle in the cytoplasm which synthesizes protein, and from the nucleus we get messenger RNA which goes to the particles. The story is a little more complicated than that, but that is its simple outline.

The mechanism is a little more complicated because we have to ask where do the small units, the aminoacids which have got to be joined together to make the protein, how do THEY get into the ribosomes. You might think possibly that the single molecules of messenger RNA could just arrange the aminoacids in the right order. But this has turned out to be too difficult a job biochemically for the nucleic acids to do. This is an example of the nucleic acids being able only to do this simple base-pairing act. What happens, therefore, is that another class of small molecules known as soluble RNA or SRNA has to be used for the function of recognition. But how do the aminoacids recognize the right sort of SRNA — and this is where the proteins come in again. The proteins join on the aminoacids to the SRNA. So in simple terms — and here I have to simplify — you have twenty sorts of aminoacids, and they are going to be joined on each one to its own sort of SRNA. At least 20 sorts of SRNA. For each aminoacid joining on to its own SRNA there is a special enzyme which can recognize these small molecules, and join them one to an other. Once the aminoacid has been joined on to the SRN'A, where it goes is determined not by itself but by the soluble RNA molecule to which it is joined. The SRNA molecule will go into the ribosome and recognize the correct sequence of basis on the messenger RNA. So the mechanism, you see, is quite complicated.

At this point I would just like to mention for the people who are interested professionally in protein synthesis a few new ideas which are partly based on experimental research but are not yet fully established. The first idea is that on any one piece of messenger RNA there are probably more than one ribosome sitting, probably 5 - 10 - 15 ribosomes perhaps on one piece of messenger RNA. The picture is as follows, if you imagine a piece of messenger RNA a ribosome will come on at one end and it will gradually work its way along to the other end. As it goes along, more and more ribosomes go on to the end and travel along, so that eventually the whole of the messenger RNA is covered with ribosomes. We do not know if this picture is correct, but it is certainly much more plausible than the previous pictures we had, and there is already considerable experimental evidence in support of something of this sort.

The second recent piece of information is the probable answer to the question: when the polypeptide chain is HALF synthesized, to what is it attached? It is almost certain that it is attached at one end, the growing end. We know that the synthesis starts at the amino- end and goes along to the carboxyl- end of the polypeptide chain. When it is half-synthesized it might be joined to the backbone of the messenger RNA. However, work in Professor Watson's laboratory and in our own laboratory suggests that the growing chain is probably joined to a molecule of soluble RNA, and it is probable that in any one ribosome there are two sites for the soluble RNA, one from the aminoacid which is just been added, and one from the next one. This is not established, but it is plausible.

Now there are many details of protein synthesis which we do not yet understand. And some of them will be very difficult to discover since they involve the stereo-chemistry of the large molecules. But without knowing these details we can ask a general question. Here in the nucleic acid we have a sequence of four things, which is somehow directing the sequence in the proteins of twenty things. This can be looked upon as a formal problem, as you might say a CODING problem. How do you translate the language of four letters in the nucleic acid into the language of twenty letters in the proteins. And this is an exciting problem because there has been recently rapid developments experimentally.

Now we can approach this in two ways - we can ask firstly general questions, and we think now we know the answers to some of these general questions. For example we can ask how many of the bases do you take at a time to stand for one aminoacid. There is now genetic evidence that suggests that you take a group of three at a time, *a triplet of bases in the nucleic acid probably stands for one aminoacid in the protein.* We cannot be sure of this on biochemical evidence here, but the genetic evidence is quite suggestive. Secondly we can ask whether if you take three bases at a time on the nucleic acid chain, do they overlap or not? We have to read these three first (Dr. Crick points to the model on the desk) and then another three starting from the last of the previous triplet so that one base will affect more than one group of three or you may simply read three and three and three: what is called non-overlapping. Well we believe on very good evidence that the code is non-overlapping, that you read the first three, numbers one, two, three, and then your read numbers four-five-six, and then you read numbers seven eight and nine. Another general question we can ask is, is the code universal? Is the relationship between the nucleic acid and the protein the same throughout nature? This question we cannot yet answer, *but we have evidence which suggests that it is similar, and it may well be the same throughout nature.* Well, there are a number of general questions which we have partial answers for.

Now we have said that it is probable that you take the bases three at a time, and you remember, there are four different sorts of them. So we can ask, how many different triplets are there? Clearly there are four times four times four, that is sixty-four possible different triplets. But we only have to code for twenty different aminoacids and possibly one or two spaces, a small number probably above twenty. So we can ask the question: it twenty four triplets are . . . . for all the to have the base uracil. But this was aminoacid and spaces, what are the remaining forty for? Are they nonsense? Or, we can ask, is it more than one triplet which stands for each aminoacid, or, in the phrase used, is the code degenerate?

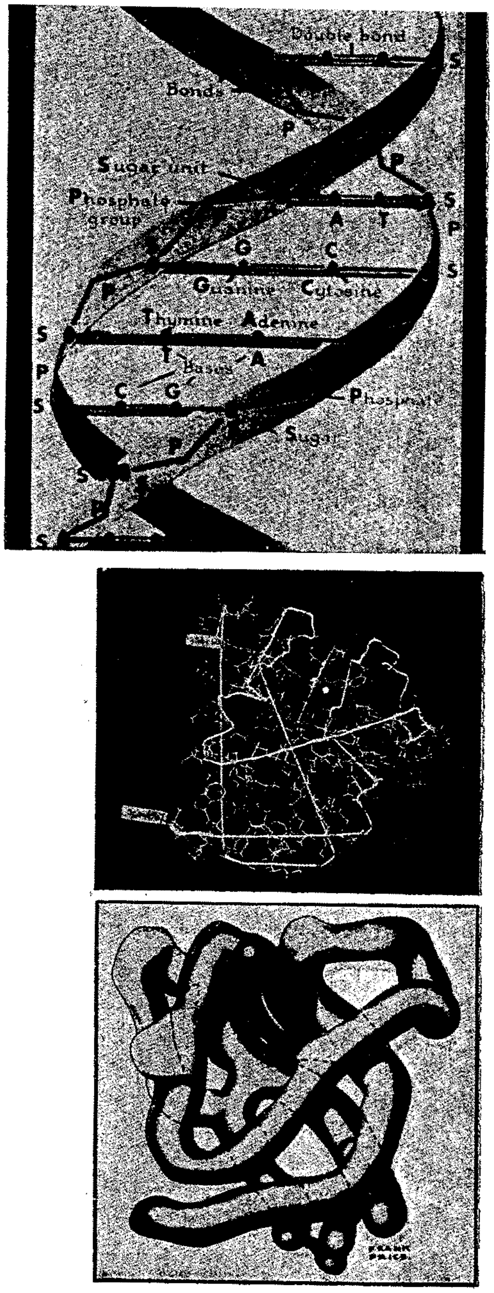
To answer this question we have to look at the different sort of work which started about a year and a half ago by Nurenberg and his colleagues who had *a test-tube system which could synthesize protein.* To this they added some specially synthesized RNA, an RNA for which every base was uracil. And they found to their great surprise that the system then produced a protein in which every aminoacid was the same, let us say, was phenylalanine. *So polyuracilic acid produced polyphenylalanine, and if the code is indeed a triplet code, this means that 3 uracils stand for phenylalanine.*

This work has been continued both by Nurenberg and by his colleagues, who have made many synthetic polynucleotides, many messenger RNAS, usually of a random sequence, and have tried them. And a number of letters of the code have . . . been produced. The work appeared to suggest that every triplet had neither probable or general grounds nor actually supported by the experimental evidence. The recent works of this group and some of our own work suggest that *in fact there are several triplets that stand for each* aminoacid. *The code is degenerate, and this is going to make it more difficult to determine which triplet stands for which aminoacid.*

However, several triplets appear to stand for one aminoacid, it is probable that the triplets which do stand for one aminoacid are rather similar. In other words, it looks as if the code has some structure, some internal relationship which we do not yet understand. In fact to a first approximation it looks almost like a doubled code, *as if two of the letters mattered and the third did not matter so much.* But this is an oversimplification. What is clear is that the code has SOME structure, but that we have not yet found it.

Let me now, having described our general ideas of nucleic acids and proteins, and how the N.A. controls the synthesis of the protein, and the problem of the code, let me now give a simple example which may be of interest to those of you with a genetical background or any of you who know it already. There is an inherited condition known as sickle-cell anemia. If you only have one of the sickle-cell genes, half your hemoglobin is abnormal. If you have sickle-cell anemia proper, then you have both genes abnormal and ALL your hemoglobin is abnormal. What is the change in the hemoglobin which you find in sickle-cell anemia? This was shown by Ingram a number of years ago in our laboratory to be *a change of just ONE aminoacid out of about three hundred*. In hemoglobin there are two chains, about a hundred and fifty aminoacids long, and out of that three hundred aminoacids only ONE aminoacid is changed. If you are so unfortunate therefore as to have this mistake in your hemoglobin, it is probable you will die before you are twenty-one.

So we must ask, in the light of



At left: article published by Watson and Crick in « Nature », 1953, which led to their award of the Nobel prize for medicine in 1962. Above: from the top, downwards: the schematic representation of DNA, a model of a highly specialized protein, myoglobin, and the same model with side chains.

our ideas, what is the change in the genetic material of the father and mother of such a person, to produce sickle-cell anemia. And it's probable that the change was just due to ONE base of the nucleic acid; one base in the father's N.A. and one base in the mother's N.A. In other words, a change of only perhaps a few atoms in the egg and sperm has produced a condition which in adult life is lethal.

But we can ask, how is it that an organism as big as ourselves can be affected by a change of a few atoms. Now we can see when we think of what happens, how this occurs. Because the DNA of the sperm and the egg which make up the genetic material of the fertilized egg is first of all copied many times, *so that you have THIS number of copies of the DNA there including the error.* Not all of those cells will make hemoglobin. But quite a large number will. They will make many copies of the messenger RNA, although this is not quite sure in hemoglobin but anyway they probably make several. And the messenger RNA which guides protein synthesis will make many copies of hemoglobin; and so you get this initial mistake, which was only a few atoms, copied many times at different levels until an enormous mass of this biological material, all the hemoglobin inside you, is defective in this way, although the original change was only to a few atoms.

The scheme I have described to you is the basic scheme which we believe controls the genetic processes and the expression of them in all living cells. Of course there are many questions of a broader type which must be answered. For example, we would like to know why it is that some genes at a particular time and others are not working. The problem of control mechanisms. This I have not touched on at all. And there are many things that are really unclear, exactly how it is that the sequence of the aminoacids determines this folding up process. Nevertheless, we believe that this scheme in its outline explains the essential properties of all biological organisms, and we believe we can go on and confirm it and extend it in detail, and from there work up to the higher levels of organization, to the embryo, embryological development, up through the cells to the tissue, and eventually to the whole organism. But at the moment, what I have described is right down at the molecular level.

But there is one idea, one central idea that I should like to leave with you, and that is the idea of the two families of molecules: *the nucleic acids, which look after replication and mutation, and the proteins, which look after the everyday work of the cells. Now we can say that life as we know it here on Earth is a symbiosis, a living together of these two families of molecules, and that an elaborate mechanism has to be made, the mechanism of protein synthesis, so that they are linked together.* Once we have seen this general thing, we understand, or we hope we understand the broad organization of living things.